Intégration de données cliniques et omiques pour la recherche d'information dans le Dossier Patient Informatisé

Chloé Cabot, Lina F. Soualmia, Stéfan J. Darmoni

Intégration de données cliniques et omiques pour la recherche d'information dans le Dossier Patient Informatisé

Chloé Cabot[1], Lina F. Soualmia[1,2], Stéfan J. Darmoni[1,2]

[1] CISMeF & TIBS, LITIS EA4108 NormaSTiC CNRS FR3638, Rouen University Hospital, France
chloe.cabot@chu-rouen.fr

[2] INSERM, LIMICS UMR 1142, Paris, France

Résumé : Nous décrivons dans cet article le modèle de données générique Information Retrieval for Omic and Clinical Sciences (IROmiCS) que nous proposons pour gérer les principaux types de données omiques (données d'expression, de méthylation de l'ADN et variants génomiques). Nous décrivons également le langage de requêtes que nous avons développé qui repose sur le modèle IROmiCS et qui est dédié à l'interrogation des données cliniques et omiques. Pour valider le modèle de données proposé, ainsi que le langage de requêtes associé, des données omiques expérimentales ont été intégrées dans ce modèle ainsi que des données de référence telle que les bases Gene du NCBI, Uniprot/Swissprot et la Gene Ontology. Plusieurs types de requêtes ciblant des données cliniques et des données omiques ont été réalisées sur les données intégrées. Une interface graphique facilite la visualisation des données intégrées par les cliniciens et les chercheurs. L'outil de recherche a permis de traiter des données symboliques, textuelles, numériques et chronologiques.

Mots-clés : Dossier Patient Informatisé, Stockage de données et Recherche d'Information, Intégration de données, Médecine personnalisée.

1 Introduction

Un Dossier Patient Informatisé (DPI) est défini comme "une version électronique du dossier papier traditionnel utilisé par les professionnels de santé" (Sewell & Thede, 2012). Hebda et Czar Hebda *et al.* (2005) décrivent le DPI comme une ressource d'information électronique utilisée en santé pour stocker les données patient. L'International Organization for Standardization (ISO) définit un DPI comme "*un entrepôt d'information sur la santé d'un individu dans une forme traitable par informatique, stockée et transmise avec une sécurité adéquate, et accessible par de multiples utilisateurs autorisés. Il respecte un modèle d'information logique communément admis, dépendant des systèmes de DPI. Son but premier est le support de la continuité, de l'efficacité et de la qualité des soins et il contient une information rétrospective, concurrente et prospective*". Selon cette dernière définition, le DPI joue un rôle central puisqu'il comporte les informations à long terme relatives aux soins et évènements de soin de tous types, mais aussi des instructions, informations prospectives comme des plans, ordres et évaluations (Garde *et al.*, 2007).

Ainsi, la communauté médicale fait face à un nouveau paradigme dans sa manière d'interagir avec les données cliniques. Les DPI permettant de gérer et partager tous les différents types de données cliniques (texte, numérique, synthétique, chronologique). Quelques entrepôts de données cliniques proposent des architectures, outils et services permettant l'utilisation des données du DPI, en particulier en recherche translationnelle. En effet, durant la dernière décennie, le séquençage de nouvelle génération (NGS) a été considéramment amélioré et ces techniques haut débit sont aujourd'hui communément utilisées pour répondre à de nombreuses questions

biologiques à l'échelle du génome : identification de variations, analyse d'expression ou encore modification de la chromatine. Alors que le génome humain avait demandé dix ans à être complété et coûté des milliards de dollars, aujourd'hui les scientifiques peuvent réaliser un séquençage de génome ou d'exome en moins d'une semaine pour moins de mille dollars (Fernald *et al.*, 2011). Les données omiques générées par l'usage croissant de ces techniques ouvrent de nouvelles perspectives dans la recherche d'applications biomédicales. Aujourd'hui, elles sont déjà utilisées pour identifier de nouveaux biomarqueurs, des mutations génétiques permettant de prédire la susceptibilité ou la prédisposition génétique à certaines pathologies ou à évaluer une réponse personnalisée à un médicament. Prochainement, la réunion de données cliniques et omiques pourrait mener à des applications innovantes comme de nouveaux tests diagnostiques ou encore des thérapies ciblées ainsi que des avancées significatives dans la compréhension de certaines maladies génétiques complexes et des mécanismes impliqués.

Plusieurs outils et frameworks dédiés à la recherche d'information dans les DPI ont été proposés. Ces outils ont été adaptés selon chaque format de données : structuré, non structuré ou mixte. Principalement deux différents types d'outils existent : (a) des Systèmes de Recherche d'Information (SRI) orientés population basés sur des entrepôts de données cliniques pour n patients et (b) des SRI au sein des DPI dédiés à un seul patient. Pour cette dernière catégorie, plusieurs outils ont été développés. CISearch (Natarajan *et al.*, 2010) a été développé et implémenté à l'hôpital universitaire de Columbia. L'utilisateur peut interroger tous les rapports textuels (imagerie, pathologie, décharge) en utilisant des outils Apache Lucène. Medical Information Retrieval System (MIRS) (Spat *et al.*, 2008) est également basé sur les outils Apache Lucène. Dans le système OpenEHR (Kalra *et al.*, 2005), un langage de requête dédié est utilisé, lié à une structure orientée archétype. Le language Archetype Query Langage (AQL) est construit pour interroger ce type de modèle de données dans les DPI. La plateforme Stanford Translational Research Integrated Database Environment (STRIDE) fournit également un système d'interrogation nommé Anonymous Cohort Tool, dédié à la création de cohortes (Lowe *et al.*, 2009). Le moteur de recherche EMERSE (Hanauer, 2006) permet la recherche plein texte avec des options avancées adaptées aux DPI comme la recherche par troncatures ou par synonymie. Enfin, XOntoRank Farfan *et al.* (2009) est un moteur de recherche sémantique permettant de réaliser des requêtes sémantiques dans des documents médicaux structurés, suivant la norme Health Level Seven International (HL7) CDA [1]. Ces documents peuvent contenir à la fois des données codées, structurées ou libres.

Integrating Biology and the Bedside (I2B2) (Murphy *et al.*, 2010) est un framework libre permettant de réutiliser les données cliniques existantes dans les DPI à des fins de recherche, et si combinées à des données génomiques, à faciliter la conception de thérapies ciblées. Cette plateforme profite actuellement d'une large adoption par la communauté scientifique académique et industrielle. L'un des composants les plus visibles d'I2B2 est le I2B2 workbench, un outil dédié à la sélection de patients permettant l'interrogation et la visualisation des données cliniques (Deshmukh *et al.*, 2009). Cependant, le modèle de données d'I2B2 n'inclut pas un point de vue centré sur le patient. Transmart (Sarkar *et al.*, 2011) est une plateforme de recherche translationnelle supportée par une communauté croissante de développeurs. Ce logiciel est directement basé sur le modèle de données d'I2B2. Il permet d'explorer des données phénotypiques, de réaliser des méta-analyses et de tester et valider de nouvelles hypothèses.

1. www.hl7.org/

Depuis 2011, un projet en cours appelé RAVEL (Recherche d'Information et Visualisation dans le Dossier Patient Informatisé) est dédié au développement d'outils efficaces et productifs permettant aux utilisateurs de situer, en temps réel, les éléments pertinents des DPI et de les visualiser grâce à des modèles de présentation synthétiques et intuitifs (Thiessard *et al.*, 2012).

Nous décrivons dans cet article le modèle de données omiques Information Retrieval for Omic and Clinical Sciences (IROmiCS) que nous avons développé. Il repose en partie sur le modèle de données cliniques RAVEL de façon à accomplir plusieurs tâches : (i) la représentation des données omiques, (ii) l'intégration, la gestion et le stockage du plus grand nombre de types de données omiques, (iii) la recherche d'information à deux échelles : la première à l'échelle d'un patient unique, qui est axée soin, et la deuxième à l'échelle de plusieurs patients, axée épidémiologie (comme par exemple la création de cohortes) et recherche clinique (comme par exemple la sélection automatique de patients pour des essais cliniques basés sur des critères d'inclusion et d'exclusion).

Cet article est organisé comme suit. La section 2 est dédiée à la description du modèle de données IROmiCS ainsi que les données qui y ont été intégrées. La section 3 présente l'outil de recherche d'information dans les DPIs via le langage de requêtes que nous avons développé, ainsi que la visualisation des données. Nous comparons nos approches avec les solutions existantes dans la section 4. Enfin, nous concluons et donnons quelques perspectives de travail dans la section 5.

2 Modèle de données IROmiCS et sources de données

2.1 Modélisation des données omiques

La conception d'un modèle de données générique gérant à la fois des données omiques et des données cliniques nécessite d'établir une revue complète et cohérente des différents types de données omiques aujourd'hui utilisés. Les données pertinentes pour l'intégration avec des données cliniques doivent être sélectionnées en fonction de leur utilité et de leur intérêt dans le cadre du DPI. Plusieurs problèmes doivent être résolus, incluant le volume de données à considérer et le manque de consensus sur les informations pertinentes à retenir.

Quatre niveaux de données ont été posés afin de décrire les types de données, en accord avec les conventions adoptées par les bases de données internationales comme ArrayExpress EBI (2013), GEO NCBI (2013) ou TCGA NIH (2013) (TABLE 1). Les données *brutes* (niveau 1) correspondent aux données non normalisées. Pour un séquençage, ce niveau correspond aux données brutes sorties du séquenceur. Elles peuvent être accessibles par des fichiers textes ou binaires, dont le format dépendra fréquemment du matériel utilisé. Le plus souvent, le volume de données est très important (jusqu'à plusieurs gigaoctets pour une seule analyse) et ces données ne peuvent pas être interprétées manuellement.

Les données *traitées* (niveau 2) correspondent aux données normalisées par une méthode statistique de régression non paramétrique comme la LOWESS par exemple. Il s'agit du signal d'une sonde ou d'un groupe de sondes pour une analyse d'expression, ou encore d'un variant supposé pour un échantillon. Ces données sont accessibles par des fichiers textes sur des banques de dépôt comme GEO ou ArrayExpress. Le volume de données est réduit, mais reste important. Pour une analyse d'expression de gènes, le fichier de résultats concernant un seul échantillon peut aller jusqu'à une centaine de mégaoctets. L'interprétation manuelle reste

Niveau	Type	Description	Exemple
1	Données brutes	Données de bas niveau par échantillon, non normalisées	Fichiers BAM ou CEL Signal brut par sonde
2	Données traitées	Données normalisées par échantillon	Signal normalisé par sonde ou set de sondes
3	Données interprétées	Données traitées agrégées par échantillon	Signal d'expression d'un gène, par échantillon
4	Régions d' intérêt	Associations quantifiées entre classes d'échantillons	Un gène X est impliqué dans 10% des lymphomes

TABLE 1 – Les quatre niveaux de données omiques permettant leur classement en fontion de leur traitement

délicate, l'information concernant des sondes ou des variants non validés.

Les données *interprétées* (niveau 3) regroupent des données qui ont été agrégées pour un échantillon. Par exemple, pour l'analyse d'expression de gènes, il s'agira du signal d'expression d'un gène, les signaux des sondes correspondant à ce gène ayant été agrégés ou encore d'un variant validé. Ce type de données est disponible en fichier texte, le plus souvent tabulé. Cependant, il n'existe pas de standard établi. Le volume de données est dans ce cas réduit, un fichier de résultats pour une analyse d'expression peut représenter de quelques kilooctets jusqu'à 1 Mo, selon le nombre de gènes analysés. Ce niveau de données n'est pas accessible dans les banques de dépôt ArrayExpress et GEO, qui ne proposent que des données de niveau 1 et 2. Peu de banques de données proposent ces données interprétées. Le portail TGCA offre les données recueillies dans une vingtaine d'études impliquant jusqu'à plusieurs centaines de patients. Les techniques utilisées sont variées et couvrent tous les types de données vus précédemment. Dans ce cas, on dispose d'informations validées et exploitables. Ce niveau de données paraît donc pertinent à intégrer dans un dossier médical. La TABLE 2 présente pour chaque type de données l'information de niveau 3 correspondante.

Enfin, les données *interprétées et agrégées* correspondent au niveau d'interprétation le plus élevé (niveau 4). Il s'agit de réaliser des associations quantifiées et croisées entre différents types d'échantillons afin d'isoler une région d'intérêt. Cette interprétation approfondie des données omiques nécessite une expertise biostatistique et biologique humaine pointue. L'aboutissement à ce niveau d'interprétation est notamment l'un des buts de la plateforme Transmart (Sarkar *et al.*, 2011). Très peu de ressources sont disponibles de façon standardisée et formalisée. De plus, de telles données ne s'appliquent plus avec l'échelle du patient, mais à celle du phénotype, il n'est donc pas adapté à l'intégration avec des données cliniques. Cependant, les régions d'intérêt isolées représentent une information pertinente, notamment à des fins de diagnostic ou de recherche.

Nous avons évalué les quatre niveaux de données afin de sélectionner les données pertinentes à modéliser. La comparaison directe de ces données et leur intégration impose de considérer certains points : (i) la normalisation des données brutes, pour exclure des biais liés à l'étude, la plateforme ou la préparation des échantillons, (ii) l'interprétation des données brutes pour améliorer la lisibilité des résultats par les cliniciens et les chercheurs et (iii) le volume de données.

Les deux premiers niveaux regroupent données brutes et traitées, qui sont de trop bas niveau et volumineuses pour être considérées. Ces données ne correspondent pas à un point de vue

centré sur un patient puisqu'elles ne sont ni agrégées ni interprétées. Cependant, le troisième niveau de données désigne des données agrégées et interprétées comme des signaux d'expression par gène par échantillon ou des variants validés. De plus, le volume de données est limité. Enfin, le quatrième niveau de données ne correspond pas avec l'échelle d'un patient puisqu'il désigne les observations faites sur une population de patients et échantillons.

Le modèle de données omique a été conçu selon le niveau 3 de données décrit. Ce modèle se compose de trois parties gérant (i) les données liées aux laboratoires et études, (ii) aux données de variants et (iii) aux données d'expression. Le détail des types de données gérées est donné dans la TABLE 2. Le modèle complet est disponible à `http://www.chu-rouen.fr/cismef/papers/omic_mld.pdf`

2.1.1 Données des études et laboratoires

La première partie du modèle vise à gérer les données des laboratoires, responsables et études. Les informations liées aux laboratoires sont leur `nom`, `code`, `adresse`, `email` et `numéro de téléphone`. Les informations stockées relatives aux études ont pour but la conservation et la traçabilité des `protocoles`, `équipements`, `échantillons` ou encore `version d'assemblage du génome` utilisé dans l'expérience ainsi que la `source des données`. Cette partie gère également les `données administratives` des responsables d'une étude.

2.1.2 Données de variants

Pour appréhender les données communément collectées liées aux variants génomiques, des collections de métadonnées de variants génomiques comme le National Center for Biomedical Ontology (NCBO) SNP Ontology [2] et la dbSNP ont été étudiées. La NCBO SNP Ontology liste 23 classes pour décrire un variant génomique, allant de la classification des acides aminés, aux données de séquençage jusqu'au type du variant. Un sous-ensemble de ces métadonnées qui peuvent être extraites des systèmes de reporting des laboratoires génomiques a été retenu. Cette partie du modèle gère ainsi des données liées aux Single Nucleotide Variants (SNV) et insertions/délétions (indels). Pour chaque variant, les `noms systématiques` (nucléiques et protéiques), les `codons` et `bases de référence et mutés`, la `catégorie` de la variation, sa `localisation` et la `région impliquée` sont retenus. Pour chaque patient, les `variations détectées` et son `génotype` pour la variation correspondante sont stockés.

2.1.3 Données d'expression, variants structuraux, méthylation de l'ADN

La base de données contient des données concernant les `gènes et protéines` extraites de la base NCBI Gene et Uniprot KB utilisées comme référence. Les analyses de méthylation de l'ADN, perte d'hétérozygotie (LOH) ou les variants du nombre de copies sont gérés grâce à une entité générique unique. Cette entité possède plusieurs attributs comme le `type de segment génomique analysé`, sa `localisation` et ses `données de référence`. Chaque gène, protéine ou segment est lié au patient concerné et au `résultat de l'analyse`.

2. `bioportal.bioontology.org/ontologies/SNPO`

Type de données	**Niveau 3 : Description**
Variants structuraux	Altération d'une région segmentée par échantillon
Analyse du nombre de copies	Altération du nombre de copies pour une région segmentée par échantillon
Méthylation de l'ADN	Valeurs bêta calculées pour une région génomique par échantillon
Expression : exon	Signal d'expression normalisé par exon par échantillon
Expression : gène	Signal d'expression normalisé par gène par échantillon
Expression : miRNA	Signal d'expression normalisé par miRNA par échantillon
Expression : jonction	Signal d'expression normalisé par jonction par échantillon
Expression : transcrit	Signal d'expression normalisé par transcrit par échantillon
Expression : protéine	Signal d'expression normalisé par protéine par échantillon
Variants (SNP, indels)	Variants validés par échantillon

TABLE 2 – Description du niveau d'interprétation 3 pour les principaux types de données omiques sélectionnés pour concevoir le modèle de données IROmiCS

2.2 Modèle de données clinique

Le modèle de données cliniques est basé sur un modèle conceptuel intégré à un modèle physique générique (Cabot *et al.*, 2014). Ce modèle conceptuel compact contient seulement une dizaine d'entités (patients, séjours, analyses et actes médicaux), alors qu'un modèle de données cliniques en comporte habituellement plus d'une centaine. Il repose sur un modèle physique générique Entité-Attribut-Valeur (EAV) composé de deux parties : le modèle définissant le modèle de données conceptuel et l'instance du modèle stockant les données. Ce modèle compact est dédié à la recherche d'information. Il permet de gérer des types de données hétérogènes. Ce "méta-modèle" intègre l'ensemble des ressources terminologiques et documents indexés. Les sources de données cliniques et omiques ont été intégrées dans le modèle de données, créant ainsi un entrepôt de données clinomiques (voir FIGURE 1).

Les données du modèle de données cliniques sont réparties en onze tables dédiées aux informations administratives du patient (table `DM_PAT`), aux analyses biologiques (table `DM_ANA`), aux prescriptions (table `DM_PRESCR`), aux actes médicaux (table `DM_ACT`), séjours (table `DM_STAY`) et comptes-rendus (table `DM_REC`). Le modèle complet est disponible à `http://www.chu-rouen.fr/cismef/papers/model_ravel.png`.

Ainsi, pour un patient donné, la base de données contiendra ses informations administratives (nom, âge, genre) et les différents séjours passés dans le centre hospitalier (avec les dates d'en-

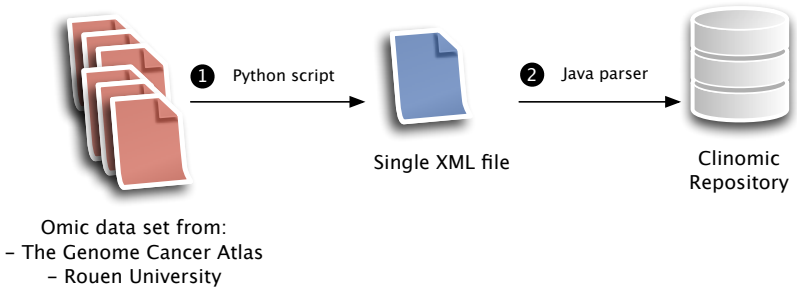

FIGURE 1 – Processus d'intégration des données omiques

Chaque jeu de données expérimentales a été traité afin de regrouper tous les fichiers de données au sein d'un fichier XML unique. Ensuite, le fichier XML a été traité et intégré dans l'entrepôt de données.

trée et de sortie et l'unité médicale d'accueil). À chaque séjour correspondra un ou plusieurs compte-rendus d'hospitalisation, ainsi que les actes médicaux et analyses biologiques réalisés ainsi que des prescriptions. Les séjours, actes médicaux, analyses biologiques, prescriptions et compte-rendus sont indexés automatiquement (Pereira *et al.*, 2009; Chebil *et al.*, 2012) par diverses terminologies (Classification Commune des Actes médicaux (CCAM), Classification Internationale des Maladies - 10^{e} édition (CIM-10), SNOMED CT, Terminologie Unifiée du Vidal).

2.3 Sources de données

Les données d'un corpus composé de 2 000 patients et 200 000 comptes-rendus ont été utilisées dans cette étude, approuvée par la Commission Nationale de l'Informatique et des Libertés (CNIL). Toutes les informations cliniques disponibles dans les DPI ont été intégrées dans le modèle RAVEL, comme les codes de la CIM10 qui permettent les codages de données comme "Cancer du colon", les données des patients (âge, genre), les résultats de tests et les comptes-rendus médicaux.

Les données omiques ont été obtenues à partir de plusieurs sources comme des bases de données internationales (Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002), ArrayExpress (Rustici *et al.*, 2013), The Cancer Genome Atlas (TCGA) (NIH, 2013)) et grâce à des collaborationss avec les laboratoires du Centre Hospitalier de Rouen (INSERM U1079 et INSERM U918) principalement spécialisés en oncologie. Les données omiques sont couplées avec des données de référence concernant les gènes, protéines et phénotypes. Pour cela, des données de référence du NCBI et Uniprot/Swissprot ont été utilisées. ces deux bases de données internationales sont supervisées et reconnues. Leurs données ont été filtrées pour ne retenir que les gènes et protéines humains dans l'entrepôt clinomique. La description des phénotypes repose sur le catalogue Online Mendelian Inheritance in Man (OMIM), et les bases Human Phenotype Ontology (HPO) (Grosjean *et al.*, 2013) et Human Rare Diseases Ontology (HRDO) (Aimé *et al.*, 2012). OMIM fournit les informations concernant les phénotypes liés aux maladies génétiques. La HRDO comporte des données sur les maladies orphelines. Enfin, la Gene Ontology complète la description des gènes et protéines. Environ (i) 9 Go de données extraites de NCBI Gene, (ii)

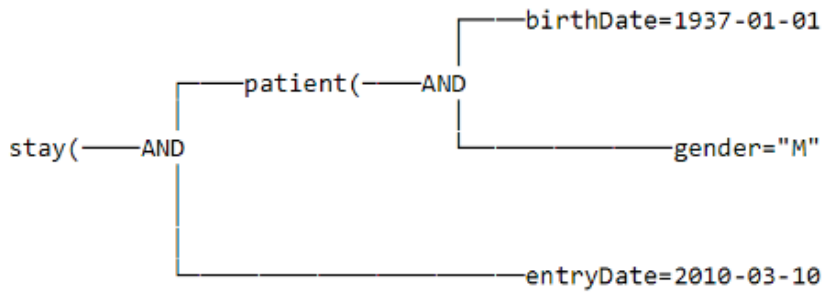

FIGURE 2 – Représentation en arbre de la requête `stay(patient(birthDate=1937-01-01 AND gender="M") AND entryDate=2010-03-10)`

530 Mo d'Uniprot/Swissprot/KB et (iii) 165 Mo d'OMIM ont été initialement intégrées dans l'entrepôt de données clinomiques. La mise à jour automatique des données est assurée quotidiennement. Pour cette étude, 88 % des maladies OMIM et 90 % des termes HPO ont été manuellement et automatiquement traduits en français et inclus dans le portail HeTOP (Grosjean *et al.*, 2013).

3 Recherche d'information

En se basant sur le modèle de données IROmiCS décrit, nous avons conçu un langage d'interrogation spécifique, dédié à la recherche d'information dans les données omiques et cliniques. Le moteur de recherche a été conçu pour être générique, rapide, multilingue et aligné avec de multiples terminologies. Il utilise un langage d'interrogation spécifique visant à faciliter la recherche d'information. Il présente trois caractéristiques principales : (i) c'est un langage orienté objet, (ii) il est flexible et (iii) il a des capacités d'interrogation complète (toutes les données contenues dans la base de données peuvent être interrogées). Le langage est composé d'unités syntaxiques, respectant la syntaxe suivante :

`ENTITÉ(CLAUSE_CONTRAINTES)`

où :

- `ENTITÉ` correspond à une entité du modèle conceptuel IROmiCS
- `CLAUSE_CONTRAINTES` correspond aux contraintes appliquées à cette entité, construite en utilisant des attributs ou des objets liés (voir TABLE 3).

Plus de détails sont disponibles sur le langage d'interrogation dans (Lelong *et al.*, 2014). Dans le modèle actuel, trois `ENTITÉS` principales sont modélisées à trois niveaux : le niveau du patient, le niveau du séjour, et le niveau le plus bas (comme l'analyse biologique ou l'analyse omique). Par exemple, les requêtes `patient()` et `medicalUnit()` retournent respectivement tous les patients et toutes les unités médicales contenues dans la base données. La clause de contrainte `CLAUSE_CONTRAINTES` permet d'appliquer des contraintes à l'entité spécifiée. C'est une expression booléenne, ainsi les opérateurs booléens `AND, OR, NOT` et les parenthèses sont utilisées pour construire des liens logiques entre contraintes uniques.

Requête en langage naturel	**Traduction dans le langage d'interrogation**

Les patients de l'étude 12 ayant une expression de HRNR supérieure à 3	`patient(`
	`study(id="OMI_STUDY_12")`
	`AND quantification(gene(geneSymbol="HRNR")`
	`AND numericValue > 3)`

Patients ayant des variations faux-sens sur HOMER1 et un taux de glucose sanguin supérieur à 1,1g/L	`patient(`
	`study(id="OMI_STUDY_1")`
	`AND variant(gene(geneSymbol="HOMER1")`
	`AND variantCategory="Missense" )`
	`AND bioTest(`
	`bioResultEXECode(label="Glucose")`
	`AND numericValue > 1.1))`

Tous les segments génomiques délétés dans l'étude 10	`quantification(interpretation="deletion"` `AND study(id="OMI_STUDY_10"))`

Tous les variants faux sens sur le gène HRNR sans l'étude 1	`variant(study(id=”OMI_STUDY_1”)`
	`AND gene(geneSymbol=”HRNR”)`
	`AND variantCategory=”Missense”)`

TABLE 3 – Exemples de requêtes omiques

Cette clause de contrainte peut être construite en utilisant les attributs de l'entité spécifiée. Par exemple, la requête suivante `patient(birthDate=1937-01-01 AND gender='M')` utilise deux attributs `birthDate` et `gender` de l'entité `patient` et retournera tous les patients masculins, nés le 01/01/1937. Les opérateurs booléens, parenthèses et comparateurs sont définis explicitement dans la grammaire du langage alors que les entités sont déduites automatiquement par auto-complétion à partir du modèle de données IROmiCS. Le moteur de recherche permet d'interpréter les requêtes pour extraire les données correspondantes de la base de données. Le processus d'interprétation contient trois étapes : (i) le parsing de la requête, (ii) sa représentation sous la forme d'un arbre (voir FIGURE 2) et (iii) la construction de la requête SQL correspondant à l'arbre généré, le modèle de données étant intégré dans une base de données relationnelle. Différentes données peuvent être extraites : (i) données symboliques (absence, présence), (ii) données numériques (avec les opérateurs >, < et =) et (iii) données chronologiques. Le moteur de recherche a été adapté pour permettre aux utilisateurs d'interroger à la fois données cliniques et données omiques (variant, gène, protéine ou segment génomique). Des mots-clés ont été définis pour interroger chaque entité du modèle omique conceptuel et élaborer des contraintes. Le temps de réponse moyen pour un patient est inférieur à deux secondes, ce qui est considéré comme satisfaisant pour un clinicien ou un chercheur. Pour n patients, le temps de réponse moyen est inférieur à dix secondes. Il est possible de réaliser la RI simultanément sur les données cliniques et omiques dans la même requête. Par exemple, les patients qui ont des variations faux-sens sur le gène HOMER1 et un taux de glucose sanguin supérieur à 1,1g/L peuvent être extraits. Chaque entité du modèle de données IROmiCS conceptuel est interrogeable. Les requêtes sont réalisables à plusieurs échelles : au niveau du patient, du séjour

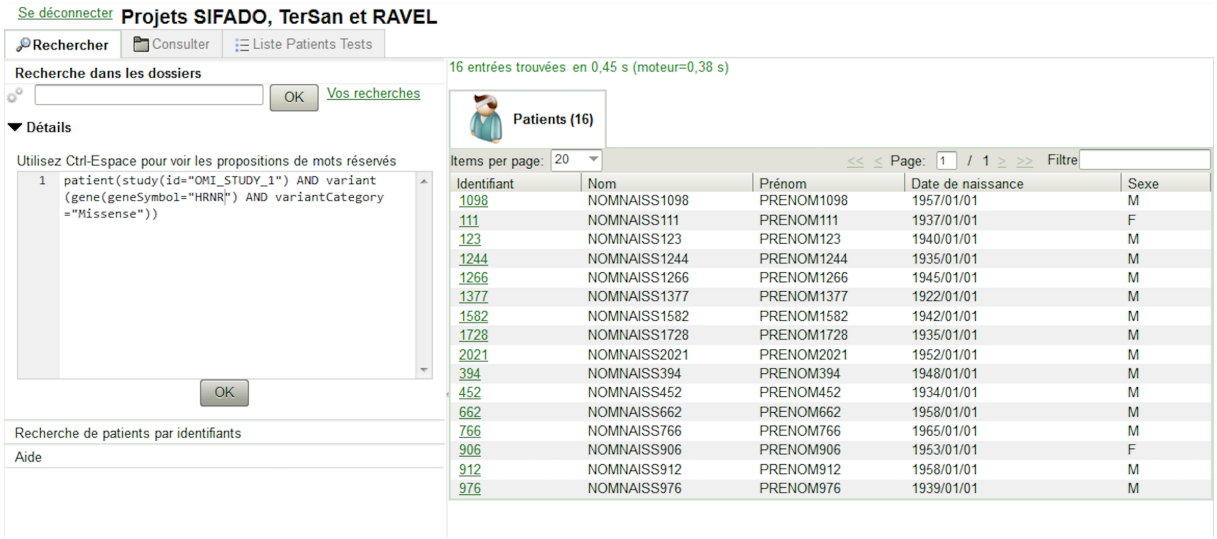

FIGURE 3 – Interface d'interrrogation

ou de l'étude. Les variants et régions génomiques peuvent également être extraits.

Pour faciliter la visualisation des données et l'utilisation de l'outil de recherche, une interface web a été conçue. Elle permet de visualiser et d'interroger toutes les données décrites dans le modèle au sein d'une interface conviviale. Les données peuvent être consultées via l'outil de recherche décrit ici et propose également une vue mono-patient où toutes les données d'un patient unique sont regroupées et visualisables (voir FIGURE 3).

4 Discussion et Conclusion

Alors que certaines solutions logicielles existent déjà en sciences translationnelles pour intégrer des données biologiques et cliniques, aucune ne gère tous les types de données omiques, données de séquences, données d'expression et variants. Bien que certaines problématiques se posent pour intégrer différents types de données omiques à partir de différentes études de sources diverses dans un même modèle de données, ce type de données peut intéresser à la fois recherche clinique et pratique clinique. Le modèle de données omique IROmiCS proposé dans cet article gère les types de données les plus communs. Il a été testé avec plusieurs jeux de données de neuf études omiques différentes. Des données d'expression (gènes, protéines, microARN), CGH-array, méthylation de l'ADN ont été intégrées avec succès. De plus, environ 25 000 variants, incluant des SNV et indels ont également été insérés avec succès dans la base de données implémentant le modèle décrit. Cependant, les variants insérés ont été extraits d'une seule étude, du fait du manque de données publiques accessibles.

Alors que la solution de référence i2b2 est largement adoptée à la fois par la communauté académique et l'industrie, notre modèle apporte certains avantages clés. En effet, IROmiCS, étendant le modèle de données clinique RAVEL peut gérer un grand nombre de types de données (numériques, dates) et est extensible et adaptable à de futurs nouveaux types de données omiques. Une interface graphique utilisateur est dédiée à la visualisation et la recherche d'in-

formation dans ces données et se base sur IROmiCS. Cette interface permet l'interrogation des données cliniques et des données omiques. De plus, le moteur de recherche développé dans le cadre du projet RAVEL peut gérer les opérateurs logiques permettant d'interroger des données numériques, et des mots clés permettant des requêtes chronologiques comme décrit précédemment. Le moteur de recherche peut gérer à la fois des requêtes multi et mono patients, alors que i2b2 ne gère que les requêtes multi-patients.

Nous envisageons d'évaluer l'ergonomie et l'utilisabilité de l'outil de recherche par un ensemble de médecins et cliniciens (avec et sans formation au langage d'interrogation). Enfin, dans le cadre de la création de cohortes, la réponse de l'outil à des critères d'inclusion et d'exclusion de patients de diverses études cliniques sera prochainement évaluée. En perspective, il pourrait être intéressant de déterminer un standard pour les données de niveaux 3, basé sur la norme HL7 RIM V3. Un tel standard serait essentiel à l'industrialisation de notre solution.

5 Remerciements

Cette étude a été financée par l'Agence Nationale de la Recherche (ANR-11-TECS-012) et la région Haute-Normandie dans le cadre du projet PLaIR2.

Références

AIMÉ X., CHARLET J., FURST F., KUNTZ P., TRICHET F. & DHOMBRES F. (2012). Rare diseases knowledge management : the contribution of proximity measurements in ontoorpha and omim. *Stud Health Technol Inform*, **180**, 88–92.

CABOT C., GROSJEAN J., LELONG R., LEFEBVRE A., LECROQ T., SOUALMIA L. & DARMONI S. (2014). Omic data modelling for information retrieval. In *Proceedings of the 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO*, p. 415–424.

CHEBIL W., SOUALMIA L., DAHAMNA B. & DARMONI S. (2012). Indexation automatique de documents en santé : évaluation et analyse de sources d'erreurs. *IRBM*, **33**(5), 316–329.

DESHMUKH V. G., MEYSTRE S. M. & MITCHELL J. A. (2009). Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC Med Res Methodol*, **9**, 70.

EBI (2013). Array express.

EDGAR R., DOMRACHEV M. & LASH A. E. (2002). Gene expression omnibus : Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**(1), 207–10.

FARFAN F., HRISTIDIS V., RANGANATHAN A. & WEINER M. (2009). Xontorank : Ontology-aware search of electronic medical records. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, p. 820–831 : IEEE.

FERNALD G. H., CAPRIOTTI E., DANESHJOU R., KARCZEWSKI K. J. & ALTMAN R. B. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics*, **27**(13), 1741–8.

GARDE S., KNAUP P., HOVENGA E. & HEARD S. (2007). Towards semantic interoperability for electronic health records. *Methods Inf Med*, **46**(3), 332–43.

GROSJEAN J., MERABTI T., SOUALMIA L. F., LETORD C., CHARLET J., ROBINSON P. N. & DARMONI S. J. (2013). Integrating the human phenotype ontology into hetop terminology-ontology server. *Stud Health Technol Inform*, **192**, 961.

HANAUER D. A. (2006). Emerse : The electronic medical record search engine. *AMIA Annu Symp Proc*, p. 941.

HEBDA T., CZAR P. & MASCARA C. (2005). *Handbook of informatics for nurses and health care professionals*. Pearson Prentice Hall.

KALRA D., BEALE T. & HEARD S. (2005). The openehr foundation. *Stud Health Technol Inform*, **115**, 153–73.

LELONG R., MERABTI T., GROSJEAN J., JOULAKIAN M., GRIFFON N., DAHAMNA B., CUGGIA M., PEREIRA S., GRABAR N., THIESSARD F., MASSARI P. & DARMONI S. (2014). Moteur de recherche sémantique au sein du dossier du patient informatisé : langage de requêtes spécifique. In *15es Journées francophones d'informatique médicale*.

LOWE H. J., FERRIS T. A., HERNANDEZ P. M. & WEBER S. C. (2009). Stride–an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc*, **2009**, 391–5.

MURPHY S. N., WEBER G., MENDIS M., GAINER V., CHUEH H. C., CHURCHILL S. & KOHANE I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*, **17**(2), 124–30.

NATARAJAN K., STEIN D., JAIN S. & ELHADAD N. (2010). An analysis of clinical queries in an electronic health record search utility. *International Journal of Medical Informatics*, **79**(7), 515–522.

NCBI (2013). Gene expression omnibus.

NIH (2013). The genome cancer atlas.

PEREIRA S., SAKJI S., NÉVÉOL A., KERGOULAY I., KERDELHUÉ G., SERROT E., JOUBERT M. & SJ D. (2009). Multi-terminology indexing for the assignment of mesh descriptors to medical abstracts in french. In *AMIA symp.*, p. 521–525 : IOS Press. PSIP.

RUSTICI G., KOLESNIKOV N., BRANDIZI M., BURDETT T., DYLAG M., EMAM I., FARNE A., HASTINGS E., ISON J., KEAYS M., KURBATOVA N., MALONE J., MANI R., MUPO A., PEDRO PEREIRA R., PILICHEVA E., RUNG J., SHARMA A., TANG Y. A., TERNENT T., TIKHONOV A., WELTER D., WILLIAMS E., BRAZMA A., PARKINSON H. & SARKANS U. (2013). Arrayexpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res*, **41**(Database issue), D987–90.

SARKAR I. N., BUTTE A. J., LUSSIER Y. A., TARCZY-HORNOCH P. & OHNO-MACHADO L. (2011). Translational bioinformatics : linking knowledge across biological and clinical realms. *J Am Med Inform Assoc*, **18**(4), 354–7.

SEWELL J. & THEDE L. (2012). Informatics and nursing : Opportunities and challenges. online glossary of terms.

SPAT S., CADONNA B., RAKOVAC I., GÜTL C., LEITNER H., STARK G. & BECK P. (2008). Enhanced information retrieval from narrative german-language clinical text documents using automated document classification. *Stud Health Technol Inform*, **136**, 473–8.

THIESSARD F., MOUGIN F., DIALLO G., JOUHET V., COSSIN S., GARCELON N., CAMPILLO B., JOUINI W., GROSJEAN J., MASSARI P., GRIFFON N., DUPUCH M., TAYALATI F., DUGAS E., BALVET A., GRABAR N., PEREIRA S., FRANDJI B., DARMONI S. & CUGGIA M. (2012). Ravel : retrieval and visualization in electronic health records. *Stud Health Technol Inform*, **180**, 194–8.